International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

A REVIEW ON MACHINE LEARNING TECHNIQUES ON SOCIAL MEDIA DATA FOR POLICY MAKING

Dr. Saloni Bhushan[1],Dr. Surabhi Shanker[2]

[1]Assistant Professor, V. K. K. Menon College, Bhandup (E), Mumbai
[2]Associate Professor, TIPS, Dwarka, GGSIPU, New Delhi

Abstract - *Social networking sites such as Twitter, Google+, Facebook and others are gaining remarkable attention in last few decades. This is contributed to the affordability of internet access and web 2.0 technologies. Social media(SM) is emerging as platform for information and opinion polls on diverse subject matters. SM is a huge data generation source. Opinion expressed in Social network can be analyzed and assist in making decision using data mining techniques. This paper aims to focus on the views and opinions of people expressed on SM about government policies and law making and how to process that real time data to get actionable insights. The main goals of governance is to provide sustainable development, security of basic rights, maximum outreach among people, SM is a very good platform to connect and study the attitude, views, and opinion of people. But given the volume of information, it's impossible to do so manually. Data needs to be extracted and preprocessed using automated tools. Machine learning algorithms can be used to identify trends and patterns in data which can be used for further course of action. This paper reviews machine learning techniques required for analyzing Big Data generated by SM to get insights that can be used for policy making.*

Key Words: *Social Media, Social Media Analysis, Big Data, Data Mining, Machine Learning, Sentiment Analysis, Opinion Mining*

1. INTRODUCTION

Affordable and omnipresent online communication on social media provides a platform for flows of ideas and opinions. It is used for information dissemination, opinion and sentiment expression, breaking news; brand marketing etc. people also use it as a forum for political debates. Government policies are also discussed and analysed on social media sites.

The main goals of governance are sustainable development, security of basic rights, maximum outreach among people, the study of people's attitude and behaviour and so on. Some opinions are positive, some are negative and some may be neutral. However, that may be a mixture of good or bad information. Some opinion may be useful to discover valuable knowledge while others may be mere assertions and therefore misleading. It needs to be understood about how online opinions emerge, diffuse, and gain momentum. Popular social media websites are Facebook, Twitter, Yahoo, Instagram, Google+ and many more. Recently Twitter became one of the most popular and reliable medium to get information.

Leveraging the power of SM, Government can have clear insights about the impact of their actions, people's opinion about some draft, their preferences and even negative reviews.

But given the volume of information, it's impossible to do so manually. The Data generated is unstructured text which is random and noisy. The challenge is to represent that raw data in suitable format to make precise prediction. Data needs to be processed so that it can be materialized for various decision and prediction using natural language processing (NLP). In the past decade, machine learning algorithms have helped to analyse historical data, often revealing trends and patterns too subtle for humans to detect. Recently, researchers are beginning to apply these algorithms to real-time data that record personal activities, conversations and movements in an attempt to improve social life.SM posts are usually a mix of text, images, sounds and videos. ML tools are perfectly adapted to such unstructured data. The reason to adopt ML in social media analysis is guided by 3 Vs of Big Data, i.e. Volume, Velocity and Variety.

The sheer volume of SM activity requires automated tools to process data. Web scrapping tools gather all the post that may be associated with a specific agenda like CAA, NRC, put them in a data lake which further fed into the algorithms to cut and process into relevant pieces. The scrapping phase relies on a keyword or by using a hash tag. More filtering levels like geo location, age , gender, speaking language can be added. Because the focus word needs to be analysed in the context it is placed in. In recent past, such types of research were done through surveys and focus group. Using ML for such purpose not only improves the accuracy, speed and reliability of answers, but it can combine different sets of pre- existing information to answer new questions. This helps in narrow down options or creating a new action course after initial testing, thus iteratively reaching a decision. ML algorithm like clustering is very much suitable in SM environment, where user sometimes mix more than one language. For example, text can be written in user's mother tongue, can have emotions which are universal and trendy hash tags in English creating a richer message that connects with global users. ML algorithms can be used to analyse different language without modifying the underlying commands. Some of the components of machine learning

include Naïve Bayes, Linear Regression, neural network and support vector machine.

2. LITERATURE REVIEW

The research on Opinion Mining started from 2003 A.D. little research had been done about people opinions and sentiments before 2003 A.D. Opinion Mining and sentiment analysis is the hottest topic in NLP. After the growth of online social networks, different research had been carried in this topic of NLP.

The different related papers and journals in the past are "**Sentiment Analysis of Twitter Data**" [1]. In this paper, tweets from the twitter are extracted and positive-negative classification is done.
The book titled "**Sentiment Analysis and Opinion Mining**"[2]is also very useful in providing the explanation of opinions and sentiments.
Similarly, the research paper entitled "**Sentiment analysis of Facebook status using Naive Bayes classifier for Language Learning**" [3] explains the use of machine learning technique Naive Bayes classifier for sentiment analysis. In this paper, social networking Facebook is used from where the posted statuses are extracted for sentiment analysis.
In 2014 A.D. research was conducted in "Public **sentiment analysis in Twitter Data for prediction of a company stock price movements**"[4]. In this paper people tweets about stock market was evaluated for prediction of fluctuation of stock price.
The research paper "**Image Popularity Prediction in Social Media using Sentiment and Context Features**"[5] deals with prediction of image popularity extracted from different social media.
Similarly, the research paper titled "**A Comprehensive survey on web content Extraction Algorithms and Techniques**" [6] explains the different machine learning algorithms used for extracting the contents of web and social media.
Another research paper entitled "**Building Lexicon for sentiment analysis from massive collection of html documents "**[7] is also equally important in explaining the sentiment analysis and opinion mining.

The research paper "**The power of prediction with social media"** [8] research paper provides an idea about prediction power of social media in marketing, stock and products. In this paper, different methods are described for popularity prediction such as sentiment index, post rate and relative strength. Similarly, different techniques such as Naive Bayes, SVM and logistic regression are also explained. Therefore, numerous researches had been performed in area of machine learning and natural processing and also going on.

3. RESEARCH METHODOLOGY

The main methodology used for this paper was through survey of journals and publications in the field of data science, machine learning and social media data management.
The data for experiment was obtained from twitter. 3000 tweets on CAA were extracted. The tweets are preprocessed and analyzed and plotted using R tool. The flow of work is as follows:

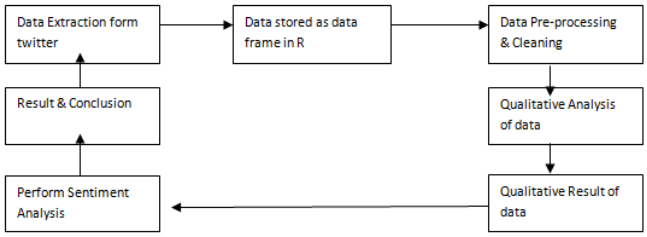

Fig 1: Research Flow Diagram

4. PREPROCESSING OF DATA

Data generated through SM sites is likely to be imperfect, containing inconsistence and redundancies. It needs to be preprocessed prior to the application of data science methods. Data preparation phase includes data transformation, integration, cleaning and normalization.

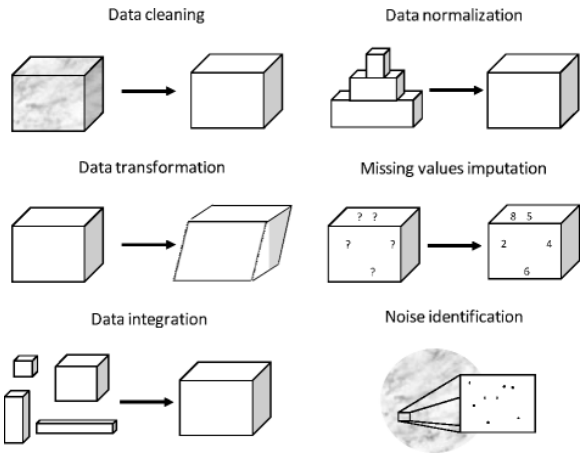

Fig 2: Data preparation phase

Data reduction phase aims to reduce the complexity of the data by feature selection, instance selection or by discretization (see Fig. 3).

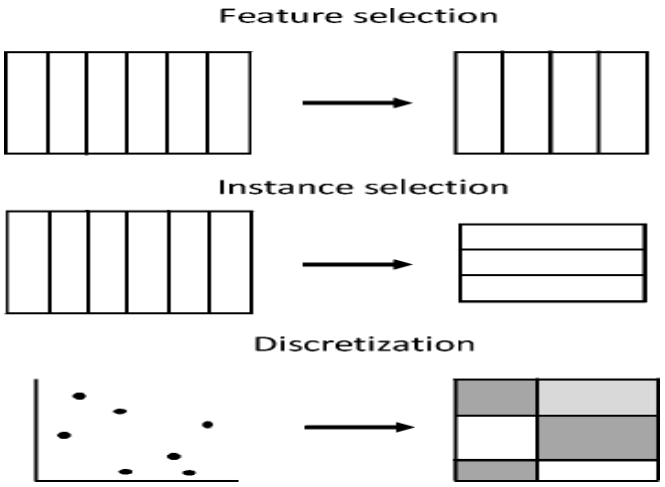

Fig 3: Data reduction phase

After the application of a successful data preprocessing stage, the final data set obtained can be regarded as a reliable and suitable source for any data mining algorithm applied afterward.

5. SENTIMENT ANALYSIS

Sentiment analysis or opinion mining is analysis of opinion or emotions from text data. It is a text analysis method that detects polarity (example Positive, negative or neutral opinion) within text, paragraph, sentence or clause posted by a person with respect to specific event. It also focus on feeling and emotions (angry, happy , sad etc) and even on intentions(example interested/not interested) .

5.1. Types of Sentiment analysis

5.1.1. Fine grained sentiment analysis

In certain situations, polarity precision is important. Polarity categories can be expanded to include very positive, positive, neutral, negative, and very negative. This is usually referred as fine grained sentiment analysis. This can be used to interpret 5 grade rating in a review. I.e.
Very positive=5; very negative=1

5.1.2. Emotion detection

This type of sentiment analysis is used to detect emotions like happiness, anger, sadness, and frustrations and so on. Emotion detection system uses lexicon, which is actually a collection of words and the emotions they conveys. But there are some limitations in using lexicon. People express emotions in different ways. Some words that typically express anger like bad or kill might also be used for expressing happiness.

5.1.3. Aspect based sentiment analysis

Sometimes, while analyzing sentiments of text, let say policy reviews, it may be desirable to know which particular aspects or feature, people are mentioning in positive, negative or neutral way. In such cases aspect based sentiment analysis can help.

5.1.4. Multilingual sentiment analysis

Multilingual sentiment analysis is a complex process. It involves lot of preprocessing and resources. Most of these lexicons are available online (example Sentiment lexicon). Others like translated corpora or noise detection algorithms needs to be created.

5.2. How sentiment analysis works

Sentiment analysis works on the basis of various Natural Language Processing (NLP) methods and algorithms. Main types are:
Rule based systems-Performs sentiment analysis based on set of manually crafted rules.
Automatic systems-Such systems rely on machine learning techniques to learn from data.
Hybrid system-Combines both rules based and automated approach.

5.2.1. Rule based approach

Generally it uses a set of human defined rules to identify subjectivity, polarity or the subject of an opinion. The rule may be based on various techniques such as
-Stemming, Tokeninzing, Part of speech tagging and parsing
-Lexicon (i.e. list of words and expressions)

5.2.2. Automatic approach

Contrary to rule based approach, it does rely on human defined rules rather depends upon machine learning techniques. A classification model can be used for sentiment analysis task, in which a classifier is fed with a text and returns its category (example positive, negative or neutral).

5.3. The training and prediction process

In training phase, the model learns to associate a particular input i.e. text to the corresponding output based on the test sample used for training. The feature extractor extracts the text input and transfers into a feature vector. Pairs of feature vector and tag (eg. Positive, negative or neutral) are fed into machine learning algorithm to build a model.

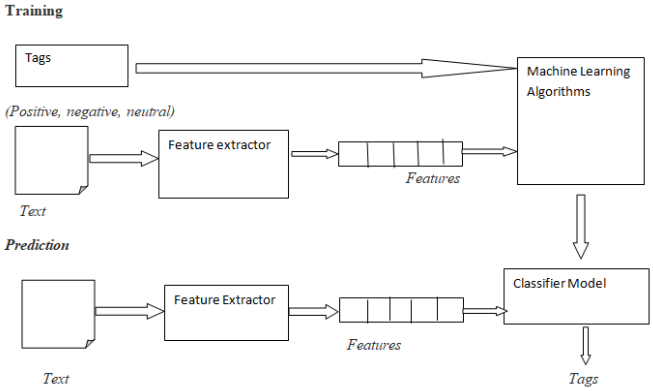

Fig 5: The training and prediction process

In the prediction phase, the feature extractor is used to transform unseen text inputs into features vectors. These features vectors are fed into the model which generates prediction tags (positive, negative or neutral) as output.

5.4. Feature Extraction

In the first step of machine learning text classifier the text is transformed into a meaningful vector of numbers. This process is known as text vectorization. The classical approach is bag-of-words or bag-of-ngrams with their frequency. Recently new feature extraction technique based on word embedding is introduced. This approach is known as word vector. In this kind of representations, words with similar meaning can be defined to have a similar representation which enhances the performance of classifiers.

5.5. Classification algorithms

Once the data is ready, next step is building a statistical model using machine learning. The data set is commonly divided into three subsets- a training set, a validation set, and a test set. The training set is used to train the model. The validation set is used to estimate how well the model is trained. The test set is used to measure the performance of model. The data set contains attributes like age, gender, relocation, language, sentiments (goal).
This step uses a statistical model like Naïve Bayes, Logistic Regression, Support Vector Machine or neural networks for classification.

5.5.1. Naïve Bayes

Probabilistic algorithms based on Bayes theorem to predict category of a text.

5.5.2. Linear Regression

Most commonly used algorithm in statistics used to predict some value(y) given a set of attributes.

5.5.3. Support vector machine

This is a non probabilistic model in which each text example is plotted as a point in n-dimensional space where n is the number of features. Different categories (sentiments) are mapped to distinct regions within the space. Then new text input are assigned a category bases on similarities with existing texts and the regions they are mapped to.

5.5.4. Deep learning

A set of diverse algorithm that attempts to mimic the human brain using artificial neural networks.

5.5.5. Hybrid approach

The desirable elements 0f rule based and automatic techniques are combined into hybrid system. The results are often more precise in this approach.

6. EXPERMENTAL WORK

Experimental work is carried out on windows platform using R studio. R is open source statistical programming language mostly used for data manipulation, data analysis, calculation and visualization of results in graphical format. It provides wide range of statistical, graphical, data mining, data analysis and sentiment analysis functionalities and excellent community support.
Present study uses R for sentiment analysis of tweeter data on Government policies like CAA. For pulling data from Twitter, Twitter developer account is required. Connecting to twitter and extracting data from Twitter is possible using R after authenticating to the Twitter API. For this experiment, 3000 sample tweets are pulled on the topic and preprocessed. Sentiment analysis algorithm is applied on preprocessed dataset to count positive, negative and neutral tweets. This experiment triesto discover how people on twitter feelabout CAA.The sentiment of a tweet are classified based on the polarity of the individual words. Each word will be given **+1**score if classified as **positive**, **-1**for**negative**, and **0**for**neutral**. To determine whether a word is positive or negative and its corresponding score, Positive and negative lexicon lists compiled in the AFINN wordlist[24] is used, which is a collection of 2477 words and phrases rated from -5 [very negative] to +5 [very positive]. AFINN words are divided into four categories:
Very negative (rating -5 or -4)
Negative (rating -3, -2, or -1)
Positive (rating 1, 2, or 3)
Very Positive (rating 4 or 5 or 6)
The total polarity score of a given tweet is calculated by adding together the scores of all the individual words in a sentence. Although polarity score is not always very accurate because it is based on isolated words rather than the overall context.

This experiment uses two packages plyr and stringr for string manipulation and ggplot2 for visualization of findings.

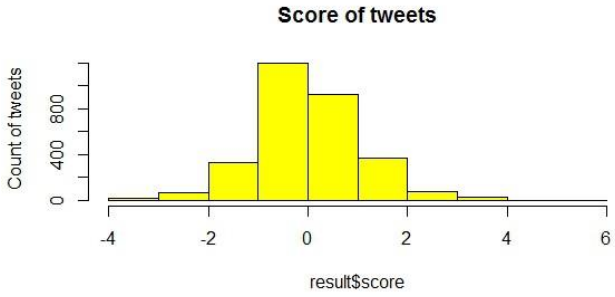

Fig 6: Histogram of the Scores

Count of tweets as per score

```
-4  -3  -2  -1   0   1   2   3
 13  41 112 531 1862 384  30  27
```

Very Negative(rating -5 or -4), Negative(rating -3,-2 or -1), Neutral(rating 0), Positive(rating 1,2,3), Very Positive(rating 1,2,3)

Count of Scores of tweets

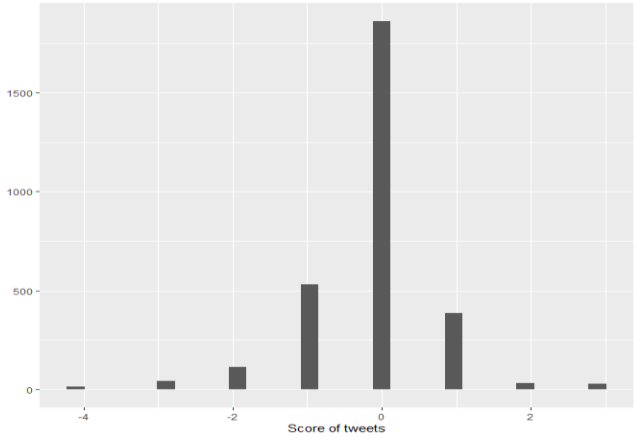

Fig 7: Count of Scores of tweets

All the tweets below 0 are considered as Negative and all the scores above 0 are considered as positive.

7. OBSERVATION

Out of 3000 tweets that were fetched from the twitter, A majority of them (1862) are neutral, whereas around (697) were having negative sentiments and around (414) tweets are positive ones.
From the plot we can see overall score is distributed between neutral, negative and positive.
However this analysis is very basic one and based only on few samples of 3000 tweets. This may vary if we increase the count of tweets as well as we apply advance filters. More advanced algorithms and tools can be used for getting a very clear view about reactions general people.

8. CONCLUSIONS

To aid the policy makers in governance various data and machine learning tools can be used. In this paper we have discussed some of the effective techniques that can be used for sentiment analysis. In future we intend to design and implement such systems for web based survey. It is useful to discover opinions of people by their tweets. It helps in identifying polarity of tweets.
Data is changing our lifestyle. The impact of Big data on our social life can not be ignored. It is affecting our lives in direct or indirect manner. The amount of data produced is increasing day by day and we definitely have to manage more data than we managing today. The better data analysis process has improved the strategy and decision making process in all sectors like governance, businesses etc, new and advanced system will have to be developed to meet future needs and machine learning algorithms are very useful in analyzing and managing big data.

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, Sentiment Analysis of Twitter Data in Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 30–38, Portland, Oregon, 23 June 2011

[2] Bing Liu, Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies ,https://doi.org/10.2200/S00416ED1V01Y201204HLT016, May 2012, 167 pages

[3] Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Jaime D. L. Caro, Sentiment analysis of Facebook statuses using Naive Bayes Classifier for language learning, Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference, July 2013

[4] Li Bing, Keith C.C. Chan, Carol Ou, Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements, 2014 IEEE 11th International Conference on e-Business Engineering, 11 December 2014

[5] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, Shih-Fu Chang, Image Popularity Prediction in Social Media Using Sentiment and Context Features, https://www.micc.unifi.it/bertini/download/papers/sp079en-gelliA.pdf

[6] Sumaia M. Al-Ghuribi, Saleh Alshomrani, A Comprehensive Survey on Web Content Extraction Algorithms and Techniques, Information Science and Applications (ICISA), 2013 International Conference, June 2013

[7] Nobuhiro Kaji, Masaru Kitsuregawa, Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), June 2007
[8] Schoen Harald , Gayo-Avello Daniel , Takis Metaxas Panagiotis , Mustafaraj Eni , Strohmaier Markus , Gloor Peter, The power of prediction with social media, https://www.emerald.com/insight/content/doi/10.1108/IntR-06-2013-0115/full/html, 14 October 2013
[9] Sergeo Consoli, Diego Reforgiato Recupero, Milan Petkovi, Data Science for Healthcare-Methodologies and Application, in ISBN 978-3-030-05248-5 and e-book ISBN 978-3-030-05249-2
[10] James A. Evans and Pedro Aceves, "Machine Translation- Mining Text for social theory", in AnnualReview of Sociology. Vol. 42:21-50
[11] H. Schoen, D. G. Avello, P. T. Metaxas, E. M. M. Strohmaier and P. Glorr, "The Power of Prediction with Social Media," in Internet Research,VOl.23 Iss: 5 pp 528-543 ,10.1108/IntR,2013.
[12] Adedoyin-Olowe, M., Gaber, M., Stahl, F.: A Methodology for Temporal Analysis of Evolving Concepts in Twitter. In: Proceedings of the 2013 ICAISC, International Conference on Artificial Intelligence and Soft Computing. 2013.
[13] Clarke, D., Lane, P., Hender, P.: Developing Robust Models for Favourability Analysis. UH Research archive,2011.
[14] García S, Luengo J, Herrera F. Data Preprocessing in Data Mining. Berlin: Springer; 2015.
[15] Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. IEEE Trans Knowl Data Eng. 2014; 26(1):97–107.
[16]https://bdataanalytics.biomedcentral.com/articles/10.1186/s41044-016-0014-0
[17] Laney D. 3D Data Management: Controlling Data Volume, Velocity and Variety. 2001.
[18]http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf..
[19]Pyle D. Data Preparation for Data Mining. San Francisco: Morgan Kaufmann Publishers Inc.; 1999.
[20]Mayra Susana Alban, David Mauricio, Predicting University Dropout through Data Mining: A Systematic Literature
[21]Adedoyin-Olowe, M., Gaber, M., Stahl, F.: A Methodology for Temporal Analysis of Evolving Concepts in Twitter. In: Proceedings of the 2013 ICAISC, International Conference on Artificial Intelligence and Soft Computing. 2013.
[22] Cathy O'Neil, Rachel Schutt "Doing data science by Oreilly"
[23] Eric Pimpler, Data Visualization and Exploration with R www2.imm.dtu.dk/pubdb/pubs/6010-full.html

BIOGRAPHIES

[1]Dr. Saloni Bhushan
Assistant Professor,
V.K.K. Menon College,
Bhandup (E), Mumbai

[2]Dr. Surabhi Shanker
Associate Professor,
TIPS, Dwarka, GGSIPU,
New Delhi